Express Mail Label No. EL426613911US 08-22-00 A

UTILITY PATENT APPLICATION TRANSMITTAL
(Large Entity)

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Docket No.
50944.8500/99RSS219

Total Pages in this Submission

JC777 U.S. PTO 08/21/00

JC987 U.S. PTO 08/22/00

TO THE ASSISTANT COMMISSIONER FOR PATENTS
Box Patent Application
Washington, D.C. 20231

Transmitted herewith for filing under 35 U.S.C. 111(a) and 37 C.F.R. 1.53(b) is a new utility patent application for an invention entitled:

METHOD FOR ROBUST CLASSIFICATION IN SPEECH CODING

and invented by:

Jes Thyssen

If a CONTINUATION APPLICATION, *check appropriate box and supply the requisite information:*

☐ **Continuation** ☐ **Divisional** ☐ **Continuation-in-part (CIP)** of prior application No.: _____

Which is a:

☐ **Continuation** ☐ **Divisional** ☐ **Continuation-in-part (CIP)** of prior application No.: _____

Which is a:

☐ **Continuation** ☐ **Divisional** ☐ **Continuation-in-part (CIP)** of prior application No.: _____

Enclosed are:

Application Elements

1. ☒ Filing fee as calculated and transmitted as described below
2. ☒ Specification having 27 pages and including the following:
   - a. ☒ Descriptive Title of the Invention
   - b. ☐ Cross References to Related Applications *(if applicable)*
   - c. ☐ Statement Regarding Federally-sponsored Research/Development *(if applicable)*
   - d. ☐ Reference to Microfiche Appendix *(if applicable)*
   - e. ☒ Background of the Invention
   - f. ☒ Brief Summary of the Invention
   - g. ☒ Brief Description of the Drawings *(if drawings filed)*
   - h. ☒ Detailed Description
   - i. ☒ Claim(s) as Classified Below
   - j. ☒ Abstract of the Disclosure

UTILITY PATENT APPLICATION TRANSMITTAL
(Large Entity)

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Docket No. 50944.8500/99RSS219
Total Pages in this Submission

Application Elements (Continued)

3. ☒ Drawing(s) *(when necessary as prescribed by 35 USC 113)*
   a. ☐ Formal Number of Sheets _____
   b. ☒ Informal Number of Sheets 4

4. ☒ Oath or Declaration
   a. ☒ Newly executed *(original or copy)* ☐ Unexecuted
   b. ☐ Copy from a prior application (37 CFR 1.63(d)) *(for continuation/divisional application only)*
   c. ☐ With Power of Attorney ☒ Without Power of Attorney
   d. ☐ *DELETION OF INVENTOR(S)*
   Signed statement attached deleting inventor(s) named in the prior application, see 37 C.F.R. 1.63(d)(2) and 1.33(b).

5. ☐ Incorporation By Reference *(usable if Box 4b is checked)*
The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied under Box 4b, is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.

6. ☐ Computer Program in Microfiche *(Appendix)*

7. ☐ Nucleotide and/or Amino Acid Sequence Submission *(if applicable, all must be included)*
   a. ☐ Paper Copy
   b. ☐ Computer Readable Copy *(identical to computer copy)*
   c. ☐ Statement Verifying Identical Paper and Computer Readable Copy

Accompanying Application Parts

8. ☒ Assignment Papers *(cover sheet & document(s))*

9. ☐ 37 CFR 3.73(B) Statement *(when there is an assignee)*

10. ☐ English Translation Document *(if applicable)*

11. ☐ Information Disclosure Statement/PTO-1449 ☐ Copies of IDS Citations

12. ☐ Preliminary Amendment

13. ☒ Acknowledgment postcard

14. ☒ Certificate of Mailing
   ☐ First Class ☒ Express Mail *(Specify Label No.):* EL426613911US

UTILITY PATENT APPLICATION TRANSMITTAL
(Large Entity)

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Docket No. 50944.8500/99RSS219
Total Pages in this Submission

Accompanying Application Parts (Continued)

15. ☐ Certified Copy of Priority Document(s) *(if foreign priority is claimed)*

16. ☐ Additional Enclosures *(please identify below):*

Fee Calculation and Transmittal

CLAIMS AS FILED

For	#Filed	#Allowed	#Extra	Rate	Fee
Total Claims	23	- 20 =	3	x $18.00	$54.00
Indep. Claims	5	- 3 =	2	x $78.00	$156.00
Multiple Dependent Claims (check if applicable) ☐					$0.00
				BASIC FEE	$690.00
OTHER FEE *(specify purpose)*					$0.00
				TOTAL FILING FEE	$900.00

☐ A check in the amount of to cover the filing fee is enclosed.

☒ The Commissioner is hereby authorized to charge and credit Deposit Account No. 19-2814 as described below. A duplicate copy of this sheet is enclosed.

- ☒ Charge the amount of **$900.00** as filing fee.
- ☒ Credit any overpayment.
- ☒ Charge any additional filing fees required under 37 C.F.R. 1.16 and 1.17.
- ☐ Charge the issue fee set in 37 C.F.R. 1.18 at the mailing of the Notice of Allowance, pursuant to 37 C.F.R. 1.311(b).

Michelle R. Orth

Signature

Dated: **August 21, 2000**

Michelle R. Orth, Reg. No. 43,944
SNELL & WILMER, L.L.P.
One Arizona Center
400 East Van Buren
Phoenix, Arizona 85004-2202
(602) 382-6275

cc:

CERTIFICATE OF MAILING BY "EXPRESS MAIL" (37 CFR 1.10) Applicant(s): **Jes Thyssen**			Docket No. **50944.8500/99RSS219**
Serial No. **TBA**	Filing Date **August 21, 2000**	Examiner **TBA**	Group Art Unit **TBA**

Invention: **METHOD FOR ROBUST CLASSIFICATION IN SPEECH CODING**

I hereby certify that this **Patent Application and accompanying paperwork**
(Identify type of correspondence)

is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 in an envelope addressed to: The Commissioner of Patents and Trademarks, Washington, D.C. 20231-0001 on **August 21, 2000**
(Date)

Suzie Mascari
(Typed or Printed Name of Person Mailing Correspondence)

Suzie Mascari
(Signature of Person Mailing Correspondence)

EL426613911US
("Express Mail" Mailing Label Number)

Note: Each paper must have its own certificate of mailing.

Title: **METHOD FOR ROBUST CLASSIFICATION IN SPEECH CODING**

Inventor: Jes Thyssen

Field of Invention

The present invention relates generally to a method for improved speech classification and, more particularly, to a method for robust speech classification in speech coding.

Background of the Invention

With respect to speech communication, background noise can include passing motorists, overhead aircraft, babble noise such as restaurant/café type noises, music, and many other audible noises. Cellular telephone technology brings the ease of communicating anywhere a wireless signal can be received and transmitted. However, the downside with the so called "cellular-age" is that phone conversations may no longer be private or in an area where communication is even feasible. For example, if a cell phone rings and the user answers it, speech communication is effectuated whether the user is in a quiet park or near a noisy jackhammer. Thus, the effects of background noise are a major concern for cellular phone users and providers.

Classification is an important tool in speech processing. Typically, the speech signal is classified into a number of different classes, for among other reasons, to place emphasis on perceptually important features of the signal during encoding. When the speech is clean or free from background noise, robust classification (i.e., low probability of misclassifying frames of speech) is more readily realized. However, as the level of background noise increases, efficiently and accurately classifying the speech becomes a problem.

In the telecommunication industry, speech is digitized and compressed per ITU (International Telecommunication Union) standards, or other standards such as wireless GSM (global system for mobile communications). There are many standards depending upon the amount of compression and application needs. It is advantageous to highly compress the signal prior to transmission because as the compression increases, the bit rate decreases. This allows more information to transfer in the same amount of bandwidth thereby saving bandwidth, power and memory. However, as the bit rate decreases, a faithful reproduction of the speech becomes increasingly more difficult. For example, for telephone application (speech signal with frequency bandwidth of around 3.3kHz) digital speech signal is typically 16 bits linear or 128 kbits/s. ITU-T standard G.711 is operating at 64 kbits/s or half of the linear PCM (pulse coding modulation) digital speech signal. The standards continue to decrease in bit rate as demands for bandwidth rise (e.g., G.726 is 32 kbits/s; G.728 is 16 kbits/s; G.729 is 8 kbits/s). A standard is currently under development that will decrease the bit rate even lower to 4 kbits/s.

Typically, speech is classified based on a set of parameters, and for those parameters, a threshold level is set for determining the appropriate class. When background noise is in the environment (e.g., additive speech and noise at the same time), the parameters derived for classification typically overlay or add due to the noise. Present solutions include estimating the level of background noise in a given environment and, depending on that level, varying the thresholds. One problem with these techniques is that the control of the thresholds adds another dimension to the classifier. This increases the complexity of adjusting the thresholds and finding an optimal setting for all noise levels is not generally practical.

For instance, a commonly derived parameter is pitch correlation, which relates to how periodic the speech is. Even in highly voiced speech, such as the vowel sound "a", when background noise is present, the periodicity appears to be much less due to the random character of the noise.

Complex algorithms are known in the art which purport to estimate parameters based on a reduced noise signal. In one such algorithm, for example, a complete noise compression algorithm is run on a noise-contaminated signal. The parameters are then estimated on the reduced noise signal. However, these algorithms are very complex and consume power and memory from the digital signal processor (DSP).

Accordingly, there is a need for a less complex method for speech classification which is useful at low bit rates. In particular, there is a need for an improved method for speech classification whereby the parameters are not influenced by the background noise.

Summary of the Invention

The present invention overcomes the problems outlined above and provides a method for improved speech communication. In particular, the present invention provides a less complex method for improved speech classification in the presence of background noise. More particularly, the present invention provides a robust method for improved speech classification in speech coding whereby the effects of the background noise on the parameters are reduced.

In accordance with one aspect of the present invention, a homogeneous set of parameters, independent of the background noise level, is obtained by estimating the parameters of the clean speech.

Brief Description of the Drawings

These and other features, aspects and advantages of the present invention will become better understood with reference to the following description, appending claims, and accompanying drawings where:

Figure 1 illustrates, in block format, a simplified depiction of the typical stages of speech processing in the prior art;

Figure 2 illustrates, in block detail, an exemplary encoding system in accordance with the present invention;

Figure 3 illustrates, in block detail, an exemplary decision logic of Figure 2; and

Figure 4 is a flow chart of an exemplary method in accordance with the present invention.

Detailed Description of Preferred Embodiments

The present invention relates to an improved method for speech classification in the presence of background noise. Although the methods for speech communication and, in particular, the methods for classification presently disclosed are particularly suited for cellular telephone communication, the invention is not so limited. For example, the method for classification of the present invention may be well suited for a variety of speech communication contexts such as the PSTN (public switched telephone network), wireless, voice over IP (internet protocol), and the like.

Unlike the prior art methods, the present invention discloses a method which represents the perceptually important features of the input signal and performs perceptual matching rather than waveform matching. It should be understood that the present invention represents a method for speech classification which may be one part

of a larger speech coding algorithm. Algorithms for speech coding are widely known in the industry. It should be appreciated that one skilled in the art will recognize that various processing steps may be performed both prior to and after the implementation of the present invention (e.g., the speech signal may be pre-processed prior to the actual speech encoding; common frame based processing; mode dependent processing; and decoding).

By way of introduction, Figure 1 broadly illustrates, in block format, the typical stages of speech processing known in the prior art. In general, the speech system 100 includes an encoder 102, transmission or storage 104 of the bit stream, and a decoder 106. Encoder 102 plays a critical role in the system, especially at very low bit rates. The pre-transmission processes are carried out in encoder 102, such as determining speech from non-speech, deriving the parameters, setting the thresholds, and classifying the speech frame. Typically, for high quality speech communication, it is important that the encoder (usually through an algorithm) consider the kind of signal and based upon the kind, process the signal accordingly. The specific functions of the encoder of the present invention will be discussed in detail below, however, in general, the encoder classifies the speech frame into any number of classes. The information contained in the class will help to further process the speech.

The encoder compresses the signal, and the resulting bit stream is transmitted 104 to the receiving end. Transmission (wireless or wireline) is the carrying of the bit stream from the sending encoder 102 to the receiving decoder 106. Alternatively, the bit stream may be temporarily stored for delayed reproduction or playback in a device such as an answering machine or voiced email, prior to decoding.

The bit stream is decoded in decoder 106 to retrieve a sample of the original speech signal. Typically, it is not realizable to retrieve a speech signal that is identical to the original signal, but with enhanced features (such as those provided by the present invention), a close sample is obtainable. To some degree, decoder 106 may be considered the inverse of encoder 102. In general, many of the functions performed by encoder 102 can also be performed in decoder 106 but in reverse.

Although not illustrated, it should be understood that speech system 100 may further include a microphone to receive a speech signal in real time. The microphone delivers the speech signal to an A/D (analog to digital) converter where the speech is converted to a digital form then delivered to encoder 102. Additionally, decoder 106 delivers the digitized signal to a D/A (digital to analog) converter where the speech is converted back to analog form and sent to a speaker.

Like the prior art, the present invention includes an encoder or similar device which includes an algorithm based on a CELP (Code Excited Linear Prediction) model. However, in order to achieve toll quality at low bit rates (e.g., 4 kbits/s) the algorithm departs somewhat from the strict waveform-matching criterion of known CELP algorithms and strives to catch the perceptually important features of the input signal. While the present invention may be but one single part of an eX-CELP (eXtended CELP) algorithm, it is helpful to broadly introduce the overall functions of the algorithm.

The input signal is analyzed according to certain features, such as, for example, degree of noise-like content, degree of spike-like content, degree of voiced content, degree of unvoiced content, evolution of magnitude spectrum, evolution of energy contour, and evolution of periodicity. This information is used to control weighting during the encoding/quantization process. The general philosophy of the present

method may be characterized as accurately representing the perceptually important features by performing perceptual matching rather than waveform matching. This is based, in part, on the assumption that at low bit rates waveform matching is not sufficiently accurate to faithfully capture all information in the input signal. The algorithm, including the present invention section, may be implemented in C-code or any other suitable computer or device language known in the industry such as assembly. While the present invention is conveniently described with respect to the eX-CELP algorithm, it should be appreciated that the method for improved speech classification herein disclosed may be but one part of an algorithm and may be used in similar known or yet to be discovered algorithms.

In one embodiment, a voice activity detection (VAD) is embedded in the encoder in order to provide information on the characteristic of the input signal. The VAD information is used to control several aspects of the encoder, including estimation of the signal to noise ratio (SNR), pitch estimation, some classification, spectral smoothing, energy smoothing, and gain normalization. In general, the VAD distinguishes between speech and non-speech input. Non-speech may include background noise, music, silence, or the like. Based on this information, some of the parameters can be estimated.

Referring now to Figure 2, an encoder 202 illustrates, in block format, the classifier 204 in accordance with one embodiment of the present invention. Classifier 204 suitably includes a parameter-deriving module 206 and a decision logic 208. Classification can be used to emphasize the perceptually important features during encoding. For example, classification can be used to apply different weight to a signal

frame. Classification does not necessarily affect the bandwidth, but it does provide information to improve the quality of the reconstructed signal at the decoder (receiving end). However, in certain embodiments it does affect the bandwidth (bit-rate) by varying also the bit-rate according to the class information and not just the encoding process. If the frame is background noise, then it may be classified as such and it may be desirable to maintain the randomness characteristic of the signal. However, if the frame is voice speech, then it may be important to keep the periodicity of the signal. Classifying the speech frame provides the remaining part of the encoder with information to enable emphasis to be placed on the important features of the signal (i.e., "weighting").

Classification is based on a set of derived parameters. In the present embodiment, classifier 204 includes a parameter-deriving module 206. Once the set of parameters is derived for a particular frame of speech, the parameters are measured either alone or in combination with other parameters by decision logic 208. The details of decision logic 208 will be discussed below, however, in general, decision logic 208 compares the parameters to a set of thresholds.

By way of example, a cellular phone user may be communicating in a particularly noisy environment. As the level of background noise increases, the derived parameters may change. The present invention proposes a method which, on the parameter level, removes the contribution due to the background noise, thereby generating a set of parameters that are invariant to the level of background noise. In other words, one embodiment of the present invention includes deriving a set of homogeneous parameters instead of having parameters that vary with the level of background noise.

This is particularly important when distinguishing between different kinds of speech, e.g. voiced speech, unvoiced speech, and onset, in the presence of background noise. To accomplish this, parameters for the noise contaminated signal are still estimated, but based on those parameters and information of the background noise, the component due to the noise contribution is removed. An estimation of the parameters of the clean signal (without noise) is obtained.

With continued reference to Figure 2, the digital speech signal is received in encoder 202 for processing. There maybe occasions when other modules within encoder 210 can suitably derive some of the parameters, rather than classifier 204 re-deriving the parameters. In particular, a pre-processed speech signal (e.g., this may include silence enhancement, high-pass filtering, and background noise attenuation), the pitch lag and correlation of the frame, and the VAD information may be used as input parameters to classifier 204. Alternatively, the digitized speech signal or a combination of both the signal and other module parameters are input to classifier 204. Based on these input parameters and/or speech signals, parameter-deriving module 206 derives a set of parameters which will be used for classifying the frame.

In one embodiment, parameter-deriving module 206 includes a basic parameter-deriving module 212, a noise component estimating module 214, a noise component removing module 216, and an optional parameter-deriving module 218. In one aspect of the present embodiment, basic parameter-deriving module 212 derives three parameters, spectral tilt, absolute maximum, and pitch correlation, which can form the basis for the classification. However, it should be recognized that significant processing and analysis of the parameters may be performed prior to the final decision. These first

few parameters are estimations of the signal having both the speech and noise component. The following description of parameter-deriving module 206 includes an example of preferred parameters, but in no way should it be construed as limiting. The examples of parameters with the accompanying equations are intended for demonstration and not necessarily as the only parameters and/or mathematical calculations available. In fact, one skilled in the art will be quite familiar with the following parameters and/or equations and may be aware of similar or equivalent substitutions which are intended to fall within the scope of the present invention.

Spectral tilt is an estimation of the first reflection coefficient four times per frame, given by:

$$\kappa(k) = \frac{\sum_{n=1}^{L-1} s_k(n) \cdot s_k(n-1)}{\sum_{n=0}^{L-1} s_k(n)^2} \qquad k = 0,1,...,3, \tag{1}$$

where $L = 80$ is the window over which the reflection coefficient may be suitably calculated and $s_k(n)$ is the k^{th} segment given by:

$$\bar{s}_k(n) = \bar{s}(k \cdot 40 - 20 + n) \cdot w_h(n), \quad n = 0,1,...79, \tag{2}$$

where $w_h(n)$ is a 80 sample Hamming window known in the industry and $s(0)$, $s(1),...,s(159)$ is the current frame of the pre-processed speech signal.

Absolute maximum is the tracking of absolute signal maximum eight estimates per frame, given by:

$$\chi(k) = \max\left\{ s(n) \mid, n = n_s(k), n_s(k)+1, ..., n_e(k)-1 \right., \qquad k = 0,1,...,7 \tag{3}$$

where $n_s(k)$ and $n_s(k)$ are the starting point and ending point, respectively, for the search of the k^{th} maximum at time $k160/8$ samples of the frame. In general, the length of the segment is 1.5 times the pitch period and the segments overlap. In this way, a smooth contour of the amplitude envelope is obtained.

Normalized standard deviation of pitch lag indicates the pitch period. For example, in voice speech the pitch period is stable, and for non-voice speech it is unstable:

$$\sigma_{L_p}(m) = \frac{1}{\mu_{L_p}(m)} \sqrt{\frac{\sum_{l=0}^{2} (L_p(m-2+l) - \mu_{L_p}(m))^2}{3}}, \tag{4}$$

where $L_p(m)$ is the input pitch lag, and $\mu_{Lp}(m)$ is the mean of the pitch lag over the past three frames, given by:

$$\mu_{L_p}(m) = \frac{1}{3} \sum_{l=0}^{2} (L_p(m-2+l). \tag{5}$$

In one embodiment, noise component estimating module 214 is controlled by the VAD. For instance, if the VAD indicates that the frame is non-speech (i.e., background noise), then the parameters defined by noise component estimating module 214 are updated. However, if the VAD indicates that the frame is speech, then module 214 is not updated. The parameters defined by the following exemplary equations are suitably estimated/sampled 8 times per frame providing a fine time resolution of the parameter space.

Running mean of the noise energy is an estimation of the energy of the noise, given by:

$$< E_{N,p}(k) >= \alpha_1 \cdot < E_{N,p}(k-1) > + (1-\alpha_1) \cdot E_p'(k), \quad (6)$$

where $E_{N,p}(k)$ is the normalized energy of the pitch period at time $k \cdot 160/8$ samples of the frame. It should be noted that the segments over which the energy is calculated may overlap since the pitch period typically exceeds 20 samples (160 samples/8).

Running mean of the spectral tilt of the noise, given by:

$$< \kappa_N(k) >= \alpha_1 \cdot < \kappa_N(k-1) > + (1-\alpha_1) \cdot \kappa(k \bmod 2). \quad (7)$$

Running mean of the absolute maximum of the noise given by:

$$< \chi_N(k) >= \alpha_1 \cdot < \chi_N(k-1) > + (1-\alpha_1) \cdot \chi(k). \quad (8)$$

Running mean of the pitch correlation of the noise given by:

$$< R_{N,p}(k) >= \alpha_1 \cdot < R_{N,p}(k-1) > + (1-\alpha_1) \cdot R_p, \quad (9)$$

where R_p is the input pitch correlation of the frame. The adaptation constant α is preferably adaptive, though a typical value is $\alpha = 0.99$.

The background noise to signal ratio may be calculated according to:

$$\gamma(k) = \sqrt{\frac{< E_{N,p}(k) >}{E_p(k)}}. \quad (10)$$

Parametric noise attenuation is suitably limited to an acceptable level, e.g., about 30 dB, i.e.

$$\gamma(k) = \{\gamma(k) > 0.968 \,?\, 0.968 : \gamma(k)\} \quad (11)$$

Noise removing module 216 applies weighting to the three basic parameters according to the following exemplary equations. The weighting removes the background noise component in the parameters by subtracting the contributions from the background noise. This provides a noise-free set of parameters (weighted parameters)

that are independent from any background noise, are more uniform, and improve the robustness of the classification in the presence of background noise.

Weighted spectral tilt is estimated by:

$$\kappa_w(k) = \kappa(k \bmod 2) - \gamma(k) \cdot <\kappa_N(k)>. \tag{12}$$

Weighted absolute maximum is estimated by:

$$\chi_w(k) = \chi(k) - \gamma(k) \cdot <\chi_N(k)>. \tag{13}$$

Weighted pitch correlation is estimated by:

$$R_{w,p}(k) = R_p - \gamma(k) \cdot <R_{N,p}(k)>. \tag{14}$$

The derived parameters may then be compared in decision logic 208. Optionally, it may be desirable to derive one or more of the following parameters depending upon the particular application. Optional module 218 includes any number of additional parameters which may be used to further aid in classifying the frame. Again, the following parameters and/or equations are merely intended as exemplary and are in no way intended as limiting.

In one embodiment, it may be desirable to estimate the evolution of the frame in accordance with one or more of the previous parameters. The evolution is an estimation over an interval of time (e.g., 8 times/frame) and is a linear approximation.

Evolution of the weighted tilt as the slope of the first order approximation, given by:

$$\partial\kappa_w(k) = \frac{\sum_{l=1}^{7} l \cdot (\kappa_w(k-7+l) - \kappa_w(k-7))}{\sum_{l=1}^{7} l^2}. \tag{15}$$

Evolution of the weighted maximum as the slope of the first order approximation, given by:

$$\partial\chi_w(k) = \frac{\sum_{l=1}^{7} l \cdot (\chi_w(k-7+l) - \chi_w(k-7))}{\sum_{l=1}^{7} l^2}. \quad (16)$$

In yet another embodiment, once the parameters of equations 6 through 16 are updated for the exemplary eight sample points of the frame, the following frame based parameters may be calculated:

Maximum weighted pitch correlation (maximum of the frame), given by:

$$R_{w,p}^{\max} = \max\{\bar{R}_{w,p}(k-7+l), l = 0,1,...,7\}. \quad (17)$$

Average weighted pitch correlation given by:

$$R_{w,p}^{avg} = \frac{1}{8}\sum_{l=0}^{7} R_{w,p}(k-7+l). \quad (18)$$

Running mean of average weighted pitch correlation, given by:

$$< R_{w,p}^{avg}(m) >= \alpha_2 \cdot < R_{w,p}^{avg}(m-1) > + (1-\alpha_2) \cdot R_{w,p}^{avg}, \quad (19)$$

where m is the frame number and $\alpha_2 = 0.75$ is an exemplary adaptation constant.

Minimum weighted spectral tilt, given by:

$$\kappa_m^{\min} = \min\{\kappa_w(k-7+l), l = 0,1,...,7\}. \quad (20)$$

Running mean of minimum weighted spectral tilt, given by:

$$< \kappa_w^{\min}(m) >= \alpha_2 \cdot < \kappa_w^{\min}(m-1) > + (1-\alpha_2) \cdot \kappa_w^{\min}. \quad (21)$$

Average weighted spectral tilt, given by:

$$\kappa_w^{avg} = \frac{1}{8}\sum_{l=0}^{7} \kappa_w(k-7+l). \quad (22)$$

Minimum slope of weighted tilt (indicates the maximum evolution in the direction of negative spectral tilt in the frame) given by:

$$\partial\kappa_w^{\min} = \min\{\partial\kappa_w(k-7+l), l=0,1,...,7\}. \tag{23}$$

Accumulated slope of weighted spectral tilt (indicates the overall consistency of the spectral evolution), given by:

$$\partial\kappa_w^{acc} = \sum_{l=0}^{7}\partial\kappa_w(k-7+l). \tag{24}$$

Maximum slope of weighted maximum, given by:

$$\partial\chi_w^{\max} = \max\{\partial\chi_w(k-7+l), l=0,1,...,7\}. \tag{25}$$

Accumulated slope of weighted maximum, given by:

$$\partial\chi_w^{acc} = \sum_{l=0}^{7}\partial\chi_w(k-7+l). \tag{26}$$

In general, the parameters given by equations 23, 25 and 26 may be used to mark whether a frame is likely to contain an onset (i.e., point where voiced speech starts). The parameters given by equations 4 and 18-22 may be used to mark whether a frame is likely to be dominated by voiced speech.

Referring now to Figure 3, decision logic 208 is illustrated in block format according to one embodiment of the present invention. Decision logic 208 is a module designed to compare all the parameters with a set of thresholds. Any number of desired parameters, illustrated generally as (1, 2, . . . k), may be compared in decision logic 208. Typically, each parameter or a group of parameters will identify a particular characteristic of the frame. For example, characteristic #1 302 may be speech vs. non-speech detection. In one embodiment, the VAD may indicate exemplary characteristic

#1. If the VAD determines the frame is speech, the speech is typically further identified as voiced (vowels) vs. unvoiced (e.g., "s"). Characteristic #2 304 may be, for example, voiced vs. unvoiced speech detection. Any number of characteristics may be included and may comprise one or more of the derived parameters. For example, generally identified characteristic #M 306 may be onset detection and may comprise derived parameters from equations 23, 25 and 26. Each characteristic may set a flag or the like to indicate the characteristic has or has not been identified.

The final decision as to which class the frame belongs is preferably decided in a final decision module 308. All of the flags are received and compared with priority, e.g., the VAD as highest priority in module 308. In the present invention, the parameters are derived from the speech itself and are free from the influence of background noise; therefore, the thresholds are typically unaffected by changing background noise. In general, a series of "if-then" statements may compare each flag or a group of flags. For example, assuming each characteristic (flag) is represented by a parameter, in one embodiment, an "if" statement may read; "if parameter 1 is less than a threshold, then place in class X." In another embodiment, the statement may read; "if parameter 1 is less than a threshold and parameter 2 is less than a threshold and so on, then place in class X." In yet another embodiment, the statement may read; "if parameter 1 times parameter 2 is less than a threshold, then place in class X." One skilled in the art can readily recognize that any number of parameters either alone or in combination can be included in an appropriate "if-then" statement. Of course, there may be equally effective methods for comparing the parameters, all of which are intended to be included in the scope of the invention.

Additionally, final decision module 308 may include an overhang. Overhang, as used herein, shall have the meaning common in the industry. In general, overhang means that the history of the signal class is considered, i.e., after certain signal classes that same signal class is favored somewhat, e.g., at a gradual transition from voiced to unvoiced the voiced class is favored somewhat in order not to classify the segments with a low degree of voiced speech as unvoiced too early.

By way of demonstration, a brief description of some exemplary classes will follow. It should be appreciated that the present invention may be used to classify speech into any number or combination of classes and the following description is included merely to introduce the reader to one possible set of classes.

The exemplary eX-CELP algorithm classifies the frame into one of 6 classes according to dominating features of the frame. The classes are labeled:

0. Silence/Background Nose
1. Noise-Like Unvoiced Speech
2. Unvoiced
3. Onset
4. Plosive, not used
5. Non-Stationary Voiced
6. Stationary Voiced

In the illustrated embodiment, class 4 is not used, thus the number of classes is 6. In order to effectively make use of the information available in the encoder, the classification module may be configured so that it does not initially distinguish between classes 5 and 6. This distinction is instead done during another module outside of the classifier where additional information may be available. Furthermore, the classification module may not initially detect class 1, but may be introduced during another module

based on additional information and the detection of noise-like unvoiced speech. Hence, in one embodiment, the classification module may distinguish between silence/background noise, unvoiced, onset, and voiced using class number 0, 2, 3 and 5 respectively.

Referring now to Figure 4, an exemplary module flow chart is illustrated in accordance with one embodiment of the present invention. The exemplary flow chart may be implemented using C code or any other suitable computer language known in the art. In general, the steps illustrated in Figure 4 are similar to the foregoing disclosure.

A digitized speech signal is input to an encoder for processing and compression into the bitstream, or a bitstream into a decoder for reconstruction (step 400). The signal (usually frame by frame) may originate, for example, from a cellular phone (wireless), the Internet (voice over IP), or a telephone (PSTN). The present system is especially suited for low bit rate applications (4 kbits/s), but may be used for other bit rates as well.

The encoder may include several modules which perform different functions. For example, a VAD may indicate whether the input signal is speech or non-speech (step 405). Non-speech typically includes background noise, music and silence. Non-speech, such as background noise, is stationary and remains stationary. Speech, on the other hand, has pitch and thus the pitch correlation varies between sounds. For example, an "s" has very low pitch correlation, but an "a" has high pitch correlation. While Figure 4 illustrates a VAD, it should be appreciated that in particular embodiments a VAD is not required. Some parameters could be derived prior to removing the noise component,

and based on those parameters it is possible to estimate whether the frame is background noise or speech. The basic parameters are derived (step 415), however it should be appreciated that some of the parameters used for encoding may be calculated in different modules within the encoder. To avoid redundancy, those parameters are not recalculated in steps 415 (or subsequent steps 425, 430) but may be used to derive further parameters or just passed on to classification. Any number of basic parameters may be derived during this step, however, by way of example, previously disclosed equations 1-5 are suitable.

The information from the VAD (or its equivalent) indicates whether the frame is speech or non-speech. If the frame is non-speech, the noise parameters (e.g., the mean of the noise parameters) may be updated (step 410). Many variations of equations for the parameters of step 410 may be derived, however, by way of example, previously disclosed equations 6-11 are suitable. The present invention discloses a method for classifying which estimates the parameters of clean speech. This is advantageous, for among other reasons, because the ever-changing background noise will not significantly affect the optimal thresholds. The noise-free set of parameters is obtained by, for example, estimating and removing the noise component of the parameters (step 425). Again by way of example, previously disclosed equations 12-14 are suitable. Based upon the previous steps, additional parameters may or may not be derived (step 430). Many variations of additional parameters may be included for consideration, but by way of example, previously disclosed equations 15-26 are suitable.

Once the desired parameters are derived, the parameters are compared against a set of predetermined thresholds (step 435). The parameters may be compared individually or in combinations with other parameters. There are many conceivable methods for comparing the parameters, however, the previously disclosed series of "if-then" statements are suitable.

It may be desirable to apply an overhang (step 440). This simply allows the classifier to favor certain classes based on the knowledge of the history of the signal. Hereby, it becomes possible to take advantage of the knowledge of how speech signals evolve on a slightly longer term. The frame is now ready to be classified (step 445) into one of many different classes depending upon the application. By way of example, the previously disclosed classes (0-6) are suitable, but are in no way intended to limit the invention's applications.

The information from the classified frame can be used to further process the speech (step 450). In one embodiment, the classification is used to apply weighting to the frame (e.g., step 450) and in another embodiment, the classification is used to determine the bit rate (not shown). For example, it is often desirable to maintain the periodicity of voiced speech (step 460), but maintain the randomness (step 465) of noise and unvoiced speech (step 455). Many other uses for the class information will become apparent to those skilled in the art. Once all the processes have been completed within the encoder, the encoder's function is over (step 470) and the bits representing the signal frame may be transmitted to a decoder for reconstruction. Alternatively, the foregoing classification process may be performed at the decoder based on the decoded parameters and/or on the reconstructed signal.

The present invention is described herein in terms of functional block components and various processing steps. It should be appreciated that such functional blocks may be realized by any number of hardware components configured to perform the specified functions. For example, the present invention may employ various integrated circuit components, e.g., memory elements, digital signal processing elements, logic elements, look-up tables, and the like, which may carry out a variety of functions under the control of one or more microprocessors or other control devices. In addition, those skilled in the art will appreciate that the present invention may be practiced in conjunction with any number of data transmission protocols and that the system described herein is merely an exemplary application for the invention.

It should be appreciated that the particular implementations shown and described herein are illustrative of the invention and its best mode and are not intended to limit the scope of the present invention in any way. Indeed, for the sake of brevity, conventional techniques for signal processing, data transmission, signaling, and network control, and other functional aspects of the systems (and components of the individual operating components of the systems) may not be described in detail herein. Furthermore, the connecting lines shown in the various figures contained herein are intended to represent exemplary functional relationships and/or physical couplings between the various elements. It should be noted that many alternative or additional functional relationships or physical connections may be present in a practical communication system.

The present invention has been described above with reference to preferred embodiments. However, those skilled in the art having read this disclosure will recognize that changes and modifications may be made to the preferred embodiments

without departing from the scope of the present invention. For example, similar forms may be added without departing from the spirit of the present invention. These and other changes or modifications are intended to be included within the scope of the present invention, as expressed in the following claims.

Abstract

A method for robust speech classification in speech coding and, in particular, for robust classification in the presence of background noise is herein provided. A noise-free set of parameters is derived, thereby reducing the adverse effects of background noise on the classification process. The speech signal is identified as speech or non-speech. A set of basic parameters is derived for the speech frame, then the noise component of the parameters is estimated and removed. If the frame is non-speech, the noise estimations are updated. All the parameters are then compared against a predetermined set of thresholds. Because the background noise has been removed from the parameters, the set of thresholds is largely unaffected by any changes in the noise. The frame is classified into any number of classes, thereby emphasizing the perceptually important features by performing perceptual matching rather than waveform matching.

Claims

1. A method for obtaining a set of parameters used for classification comprising the steps of:

(a) receiving a signal at a processing unit;

(b) providing at least one basic parameter corresponding to the signal;

(c) if present, estimating a noise component of the parameter; and

(d) if present, removing the noise component from the parameter.

2. The method of claim 1 further comprising the step of determining whether the signal is speech or non-speech.

3. The method of claim 1 further comprising the step of providing at least one additional parameter.

4. The method of claim 3 wherein the noise component is present and the step of providing at least one additional parameter is in response to the noise component.

5. The method of claim 2 further comprising the step of updating the noise parameters if the signal is non-speech.

6. The method of claim 1 wherein the step of providing comprises deriving at least one basic parameter corresponding to the signal.

7. The method of claim 1 wherein the step of providing comprises receiving at least one basic parameter corresponding to the signal.

8. A method for classifying speech comprising the steps of:

(a) receiving a speech-related signal at a processing unit;

(b) providing at least one parameter to be used for classifying the signal;

(c) estimating a noise component of the parameter;

(d) removing the noise component from the parameter;

(e) comparing the parameter with a set of at least one threshold; and

(f) associating the signal with a class in response to the comparing step.

9. The method of claim 8 further comprising the step of determining whether the signal is speech or non-speech.

10. The method of claim 9 further comprising the step of updating a noise component if the signal is non-speech.

11. The method of claim 8 wherein at least one parameter is derived to classify the signal.

12. The method of claim 11 wherein a set of basic parameters is derived and at least one noise component parameter.

13. The method of claim 8 wherein said comparing step comprises:

(a) identifying at least one characteristic of the signal with at least one the parameters;

(b) setting a flag to indicate the characteristic is present;

(c) receiving at least one flag in a final decision module; and

(d) associating a class with at least one flag.

14. The method of claim 8 wherein at least one parameter is received to classify the signal.

15. A method for perceptually matching a speech signal in a speech coding device having at least one process module, the method comprising the steps of:

(a) receiving the signal at the speech coding device;

(b) deriving a plurality of signal parameters in the process module;

(c) weighting the parameters;

(d) associating a particular signal characteristic with the signal parameters;

(e) setting a flag in the process module when the characteristic is identified;

(f) comparing the flags; and

(g) classifying the signal according to one of the comparing step or the deriving step.

16. The method of claim 15 wherein said deriving step comprises deriving a set of basic parameters and deriving a set of noise-related parameters.

17. The method of claim 15 wherein said weighting step comprises:

(a) estimating a noise component of the parameter in the process modules; and

(b) removing the noise component of the parameter in the process module.

18. The method of claim 17 wherein said weighting step comprises a set of noise estimation equations.

19. A method for speech coding whereby a set of homogeneous parameters is provided for classifying a signal, the set of parameters being uninfluenced by a background noise.

20. A method for speech communication whereby influence from speech-related noise is reduced, the method comprising the steps of:

(a) receiving a digital speech-related signal at a speech processing device;

(b) forming a set of homogenous parameters;

(c) comparing the parameters with a threshold; and

(d) classifying the signal.

21. The method of claim 20, wherein the forming step comprises forming a set of "noise-free" parameters.

22. The method of claim 21, wherein the forming step comprises:

(b1) estimating a noise component; and

(b2) removing the noise component.

23. The method of claim 20, wherein the comparing step is with a set of thresholds.

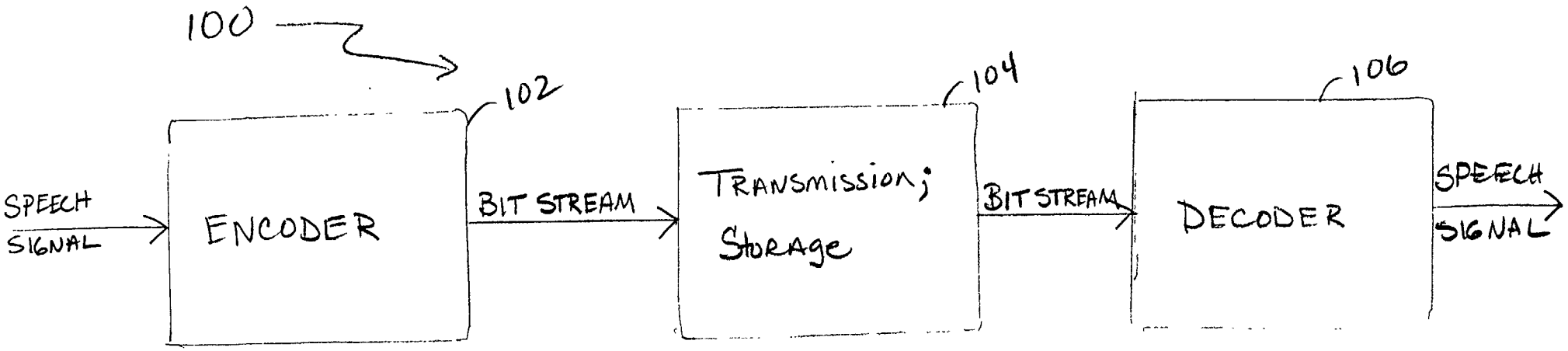

Figure 1
Prior Art

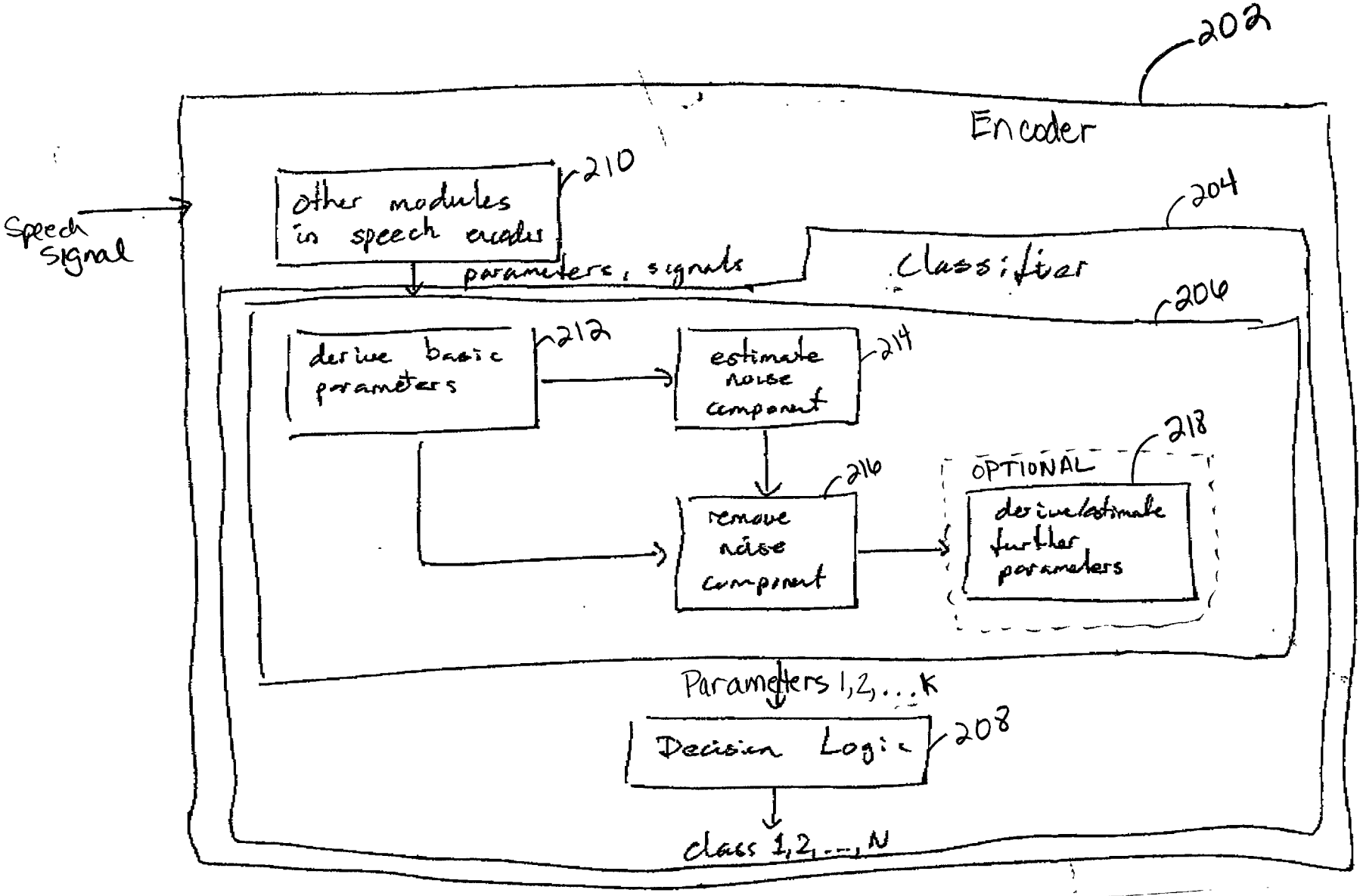

Figure 2

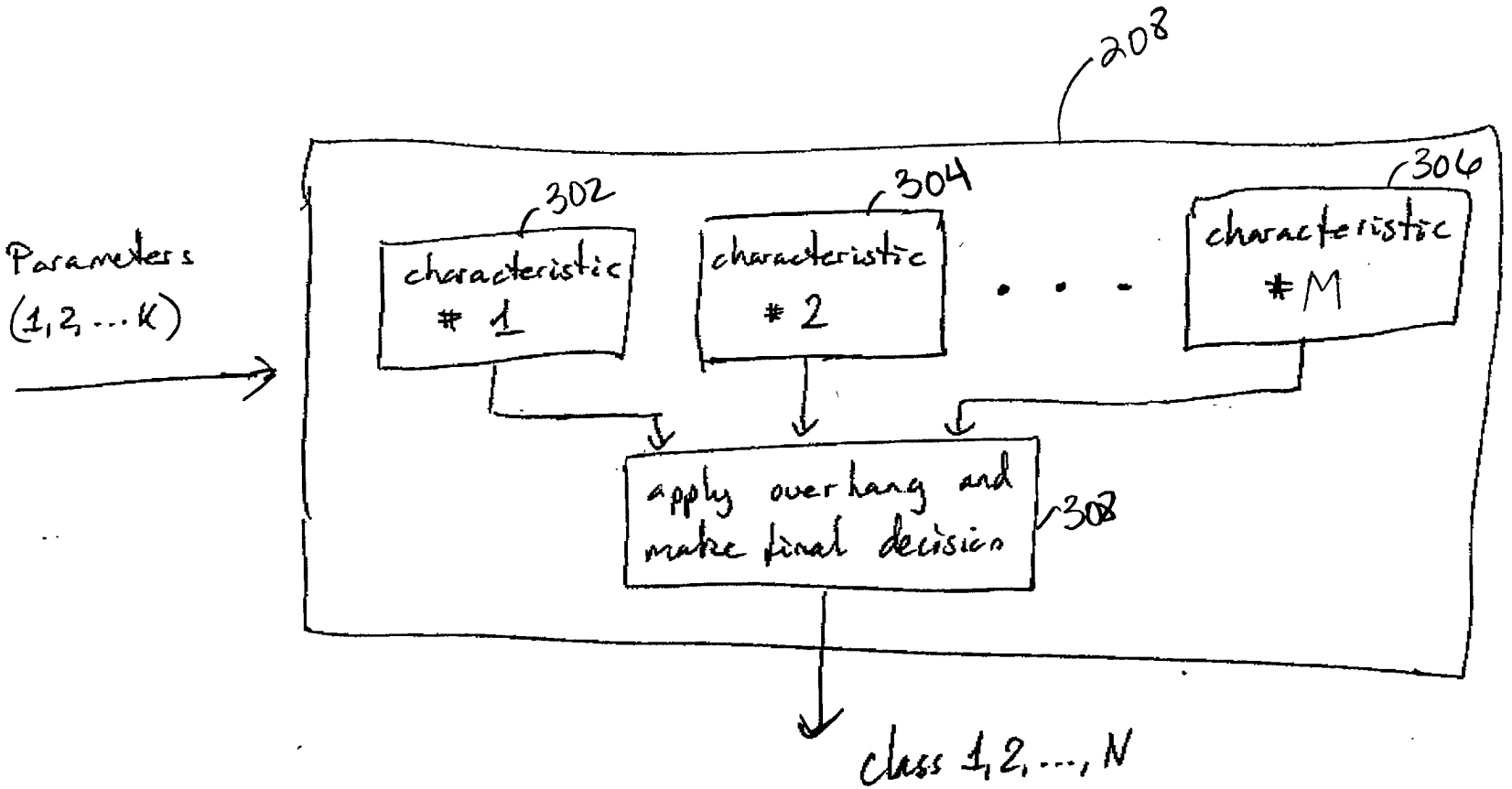

Figure 3

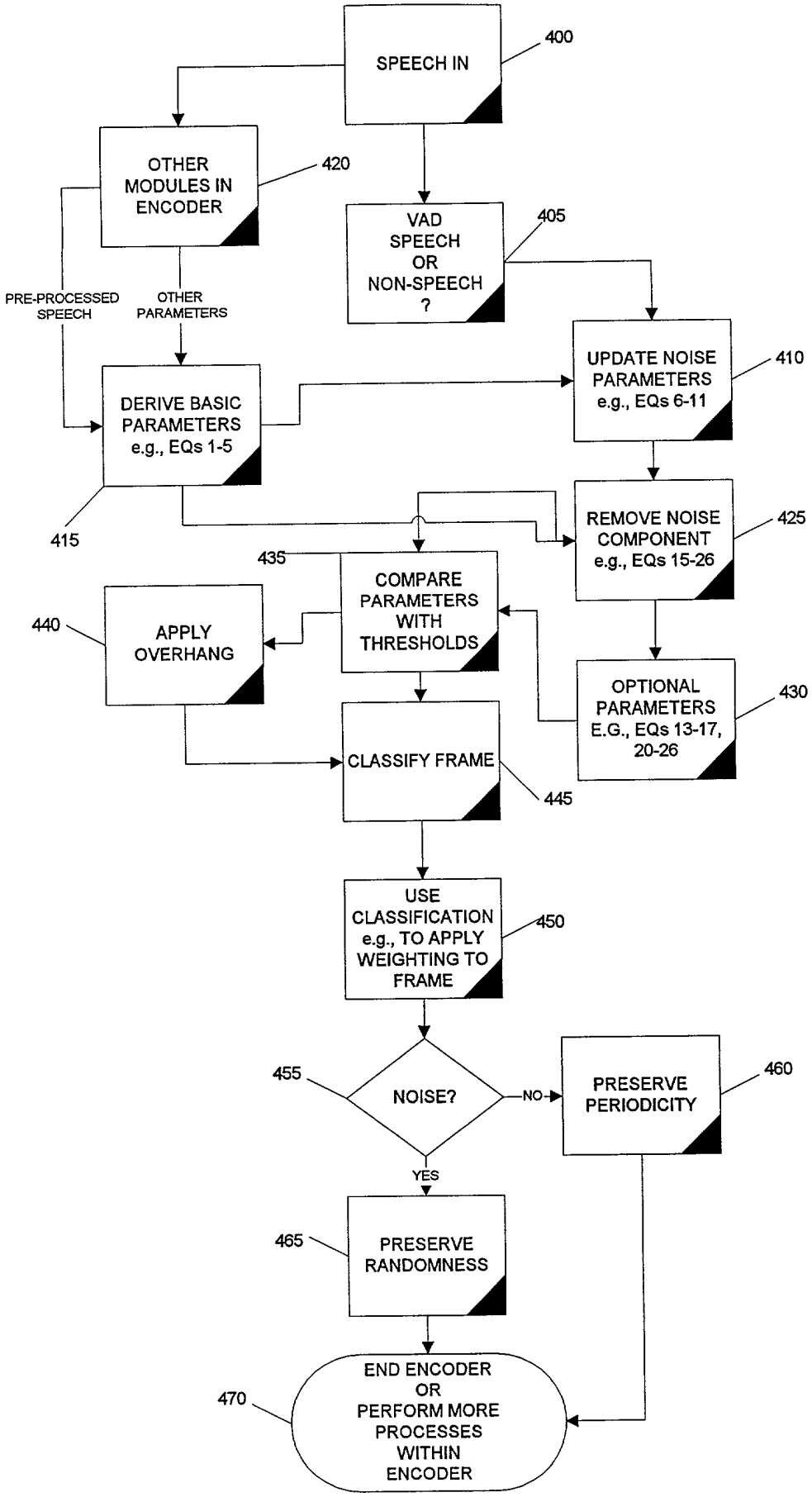
SPEECH IN
400
OTHER MODULES IN ENCODER
420
VAD SPEECH OR NON-SPEECH ?
405
PRE-PROCESSED SPEECH
OTHER PARAMETERS
UPDATE NOISE PARAMETERS e.g., EQs 6-11
410
DERIVE BASIC PARAMETERS e.g., EQs 1-5
415
REMOVE NOISE COMPONENT e.g., EQs 15-26
425
435
COMPARE PARAMETERS WITH THRESHOLDS
440
APPLY OVERHANG
OPTIONAL PARAMETERS E.G., EQs 13-17, 20-26
430
CLASSIFY FRAME
445
USE CLASSIFICATION e.g., TO APPLY WEIGHTING TO FRAME
450
455
NOISE?
NO
PRESERVE PERIODICITY
460
YES
465
PRESERVE RANDOMNESS
470
END ENCODER OR PERFORM MORE PROCESSES WITHIN ENCODER

Figure 4

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s):	Jes Thyssen	Atty Docket No.:	50944.8500
Serial No.:		Client Ref:	99RSS219
Filed:		Group Art Unit:	
TITLE:	METHOD FOR ROBUST CLASSIFICATION IN SPEECH CODING	Examiner:	NYA

DECLARATION FOR PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name.

I believe I am an original, first and joint inventor of the subject matter which is claimed and for which a patent is sought on the invention entitled METHOD FOR ROBUST CLASSIFICATION IN SPEECH CODING, the specification of which:

[X] is attached hereto.
[ ] was filed on _____ as Application Serial No. _____ and was amended on _____ (if applicable).

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to patentability as defined in 37 C.F.R. §1.56.

I hereby claim foreign priority benefits under 35 U.S.C. § 119(a)-(d) or § 365(b) of any foreign application(s) for patent or inventor's certificate, or § 365(a) of any PCT International application which designated at least one country other than the United States, listed below and have also identified below, by checking the box, any foreign application for patent or inventor's certificate, or PCT International application having a filing date before that of the application on which priority is claimed.

			Priority Not Claimed
None			[ ]
Number	Country	Filing Date	
			[ ]
Number	Country	Filing Date	

I hereby claim the benefit under 35 U.S.C. § 119(e) of any United States provisional application(s) listed below.

None	
Application Number	Filing Date
Application Number	Filing Date

I hereby claim the benefit under 35 U.S.C. §120 of any United States application(s), or §365(c) of any PCT International application designating the United States, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of 35 U.S.C. §112, I acknowledge the duty to disclose information material to patentability as defined in 37 C.F.R. §1.56 which occurred between the filing date of the prior application and the national or PCT international filing date of this application.

None		
Application Serial No.	Filing Date	Status — Patent, Pending, Abandoned
Application Serial No.	Filing date	Status — Patent, Pending, Abandoned

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full name of first inventor: Jes Thyssen

Inventor's signature: Jes Thyssen

Date: Aug. 21, 2000

Residence: Laguna Niguel (City) CA / USA (State/Country)

Citizenship: Denmark

Post Office Address: 96 Chandon Zip Code: 92677